Computers and Electrical Engineering 000 (2017) 1–13

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Employing fisher discriminant analysis for Arabic text classification☆

Dia AbuZeina, Fawaz S. Al-Anzi*

Department of Computer Engineering, Kuwait University, Kuwait City, Kuwait

ARTICLE INFO

Article history:
Received 19 March 2017
Revised 29 October 2017
Accepted 3 November 2017
Available online xxx

Keywords:
Arabic
Text
Classification
Linear discriminant analysis
Eigenvectors
Fisher

ABSTRACT

Fisher's discriminant analysis; also called linear discriminant analysis (LDA), is a popular dimensionality reduction technique that is widely used for features extraction. LDA aims at finding an optimal linear transformation based on maximizing a class separability. Even though LDA shows useful results in various pattern recognition problems, such as face recognition, less attention has been devoted to employing this technique in Arabic information retrieval tasks. In particular, the sizable feature vectors in textual data enforces to implement dimensionality reduction techniques such as LDA. In this paper, we empirically investigated an LDA based method for Arabic text classification. We used a corpus that contains 2,000 documents belonging to five categories. The experimental results showed that the performance of semantic loss LDA based method was almost the same as the semantic rich singular value decomposition (SVD), and that is indication that LDA is a promising method for text mining applications.

1. Introduction

Linear discriminant analysis (LDA), also known as Fisher's LDA, is one of the popular dimensionality reduction techniques that can show good performance in pattern recognition tasks. LDA is a classical method for projection of high-dimensional feature vectors into a lower dimensional through a linear transformation process that best separates the data groups. The primary goal of LDA is to find the best linear approximations of object feature vectors for effective and reasonable use in various classification tasks. Reference [1] demonstrates two major objectives in the separation of groups, the description of group separation, and prediction or allocation of observations to groups. In general, dimensionality reduction techniques aim at producing a compact data set given the original incompact data set. Unlike principal component analysis (PCA) [2] that works in unsupervised mode, the LDA is a supervised technique that requires class labels of the training data. The new transformed space is obtained using eigenvalues decomposition of both the between-class covariance matrix and the within-class covariance matrix. Both LDA and PCA are well-known methods that have a distinction in which the LDA focuses on discrimination while PCA focuses on representation.

The literature shows that dimensionality reduction has several advantages, such as facilitating data visualization and interpretation. It also can help in coping with the curse-of-dimensionality problem. Nevertheless, minimizing feature vectors should always preserve the essential information of the original data without sacrificing the classification performance. That is, the loss of information in the new feature space is as small as possible. Besides the LAD and PCA, there are some other

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. S. A. Aljawarneh.

* Corresponding author.
E-mail addresses: dia.abuzeina@ku.edu.kw (D. AbuZeina), fawaz.alanzi@ku.edu.kw, alanzif@eng.kuniv.edu.kw (F.S. Al-Anzi).

https://doi.org/10.1016/j.compeleceng.2017.11.002
0045-7906/

Table 1
Classification and features related methods.

Classification methods	Features selection
Euclidian distance , cosine similarity measure, Naïve Bayes (NB), *k*-Nearest Neighbors (k-NN), Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), classification tree (CT), induction rule (CN2), Rocchio, Dice distance, Maximum Entropy (ME).	Document frequency (DF), information gain (IG), mutual information (MI), Chi-Squared (χ^2-text = CHI), term strength (TS), N-grams, stems, roots. **Features Extraction** LDA, PCI, SVD

popular dimensionality reduction techniques such as singular value decomposition (SVD) [3] and Laplacian eigenmaps [4]. Intuitively, no single algorithm can be best suited for all applications. The algorithm's selection is based on the reduction time, improved accuracy, and simplified representation. For classification, a distance metric such as Euclidean distance, Mahalanobis distance, or Cosine similarity is used along with the new features to assign the instance in question to one of the training groups.

No doubt, large textual feature vectors characterize text classification problems. The vector space model (VSM) is a popular textual representation method. However, VSM is used the entire vocabulary in order to represent a single document. Hence, intelligent methods are required to minimize the feature dimensions for effective use in various machine-learning techniques. In fact, employing features reduction is, indeed, the key to the performance of the classifiers as well as cheaper in computation. Finding a linear good projection is extremely important especially in text mining applications that usually have a large number of words, sometimes beyond the hardware capabilities. In most cases, text classification tasks are of ten thousand dimensions in common; therefore, a dimensionality reduction technique can be thorough as a data compression technique since it projects the original data into a desired dimensionality of the smaller space. Nevertheless, LDA might fail to find good projection if the instances of the training classes are not well separated. Still LDA has some drawbacks, such as the singularity problem and the semantic loss of the generated features. In contrast, latent semantic indexing (LSI) employs an SVD method to generate semantic rich features. Semantic rich means that the method preserves and understands the intrinsic latent relationships between the words in the different documents.

The VSM challenges such as high dimensions and sparsity pose to carefully consider the textual features. In general, there are two methodologies for handling features; features selection and feature extraction. Feature selection is the process of selecting a subset of "hopefully" discriminative features or the most important features form the original features according to some relevant criterion. For instance, document frequency (DF) is a feature selection method that can be used to select the words that appear in a particular number of documents. In this case, the selected features are not physically changed compared to the features extraction process in which new features sets are generated. That is, the feature extraction methods focus on building a new feature set based on the original features using some techniques such as LDA, PCA, and SVD. Table 1 shows some of the popular text classification and features related algorithms. As shown in the table, there are many machine-learning options to perform text classification. In this paper, we used, as highlighted in the table, Euclidean distance for classification and DF for features selection. For feature extraction, we used LDA method. In fact, the simple Euclidian distance classifier was used, as our focus is to find the capacity of LDA feature extraction method rather than exploring the classification techniques.

The features extraction methods shown in Table 1 are different in some aspects. For instances, LDA assumes labeled training data whereas PCA assumes unlabeled data (i.e. ignores class labels). In addition, SVD is a semantic support method while both LDA and PCA are semantic loss techniques. There is a large research record focusing on the techniques to extract the optimal textual features. However, the methods that use semantic rich features are of growing interest. As a common property, these methods use Eigen values and Eigen vectors as a base for finding the important features. The information that is provided in table highlights the main directions in text classification field.

This paper aims at exploring the capabilities of the LDA for Arabic text classification task. The LDA initially was introduced for two-class classification problems. Later, the extension of the two-class LDA was known as multiple discriminant analysis that is used for multiclass classification tasks. Hence, we used LDA for multiclass text classification. In this work, we discuss two multiclass LDA based methods known as generalized eigenvectors of the ratio (inverse within-class and between-class scatters) and linear classification functions where each class has its linear classification rule. We used an Arabic corpus that contains 2000 documents, 1750 documents for training and the rest is for testing. The Matlab tool was used for both the LDA process and the classification process.

The rest of the paper is organized as follows: in the next section, we present some of the Arabic text classification challenges. Section 3 presents the literate review of LDA followed by LDA background in Section 4. Section 5 demonstrates the proposed method and the experimental results in Section 6. Finally, we conclude in Section 7.

2. Text classification challenges

The VSM model has two well-known problems; which are high dimensionality and sparsity of the data. Moreover, Text classification has other challenges, such as noises and mixed contents documents. Noises are the words that have little classification capabilities (i.e. no discriminative power) such as small words prepositions {من, الى, في, عن} <>{from, to, in,

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

An example of Arabic article

في امسية احتفالية واجواء طغى عليها الطابع الرمضاني وتخللتها العديد من الفقرات والجوائز القيمة اقام البنك الاهلي الكويتي غبقته السنوية وذلك يوم الثلاثاء الموافق يونيو بقاعة الراية التابعة لفندق ماريوت كورت يارد شهدت الامسية التي ضمت اكثر من موظف من مختلف الادارات العديد من الفقرات والفعاليات الممتعة والجوائز القيمة التي فاز بها الموظفون في البداية رحب الرئيس التنفيذي للبنك الاهلي الكويتي ميشال العقاد بالموظفين موجها لهم كلمة بمناسبة حلول شهر رمضان الكريم وشكر الموظفين على ما يبذلونه من جهد لتقديم افضل الخدمات المصرفية متمنيا لهم الاستمرار في العطاء والعمل الجاد تضمنت الامسية العديد من الفقرات والسحوبات التي منحت الموظفين فرصة الحصول على جوائز قيمة من بينها هواتف سامسونغ وشاشات سامسونغ وتذاكر سفر الى لندن وباريس واسطنبول ودبي بالاضافة الى الجائزة الكبرى وهي سيارة جي ام سي تيرين الجديدة والتي كانت من نصيب الفائز اسرار احمد خان وقد اضاف لأجواء الامسية وجود اثنين من فناني الكاريكاتير وهما السيد محمد القحطاني والسيد محمد خليل اللذان امتعا الحضور برسوماتهم الكاريكاتيرية المميزة للموظفين طوال الامسية كما استمتع الحضور بالموسيقى على انغام الفرقة الكويتية وذلك وسط اجواء رمضانية مبهجة

The translation using Google translator

In the evening festive atmosphere overshadowed by the character of Ramadan and punctuated by several paragraphs and prizes hosted Ahli Bank of Kuwait Annual Gbakth and on Tuesday, June Hall of the Hotel Courtyard Marriott banner saw the evening, which included most of the employees from the various departments of several paragraphs and events fun and valuable prizes that won by staff at first Chief Executive of the Bank of Kuwait Ahli Michel Accad welcomed staff directed them to a speech marking the holy month of Ramadan and thanked the staff for their efforts to provide the best banking services and wished them to continue in the tender and hard work included the evening several paragraphs and raffles that gave employees a chance get valuable prizes including Samsung and screens, Samsung and tickets to travel to London, Paris, Istanbul and Dubai phones in addition to the grand prize, a GMC Terrain's new car, which went to the winner of the secrets Ahmed Khan has been added to the atmosphere of the evening and the presence of two artists, cartoonists, namely Mr. Mohammad Al-Qahtani Mr. Mohamed Khalil, who Amtaa attendance Brsumaathm cartoons special staff throughout the evening as attendees enjoy music on the music of the Kuwaiti band amid the atmosphere of Ramadan exhilarating

Fig. 1. Religion article with translation using Google translator.

for}. In general, such noises are generally handled by adding these types words to the stop word list to be ignored before the classification process. However, the true challenge in text classification is the mixed contents documents that have words that belong to completely different categories. In this case, the classification process might give less than optimal performance and, therefore, increases the misclassification rate. For instance, Fig. 1 is an article that belongs to the Religion category, based on our training data; however, it has some words that belong to other categories.

Based on the Fig. 1, the following are some instances of the words that belong to different categories. The article has words of the Technology category such as: (غوغل=Samsung, تاشاش=screens, فتاوه=phones), the Economy category: (البنك =bank, الرئيس التنفيذي=chief executive, قرايس=car, الخدمات المصرفية=banking), the Tourism category: (رفس راكاذت= traveling tickets,لندن =London, باريس=Paris, دبي=Dubai, فندق = hotel, اسطنبول= Istanbul), the Sport category: (فائز = winner, جائزة = prize), and the Art category (فنان = artist, موسيقى = music, الرسومات الكاريكاتيرية=cartoons). In this article, only one word is directly related to the Religion category that is (رمضان= Ramadan). However, this word might be a person's name. In addition, other words are neither stop word nor are the other categories' words, such as the word (اسرار) which means (secrets), as it is used in this article as a person's name. The Arabic article in Fig. 1 was translated into English using the Google Translator available in [5]. The document intentionally has no numbers, commas, full stops, English characters, or other symbols such as :{à ® é ² è ² ™ ¯ > × — ô é ü ¹ × ‘ ° © ½ ¾ ë ’ ” ; ؛ –- •% ’ * , : ! ? ؟} due to the fact that the documents have to exclusively contain Arabic letters as the preprocessing stage is generally the first step in text classification.

In fact, the word is the basic unit in text classification. Hence, if the documents contain different words that belong to different categories as indicated in Fig. 1, errors might be increased that reduce the overall performance. Indeed, text classification requires efficient methods for choosing the high quality discriminative words especially when the corpus's documents show content diversity. For more details of Arabic challenge, the reader refers to [6]. For further illustration,

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

An example of Arabic article

التداوي باللصقات التي تحتوي على صفائح الذهب
ما الحكم الشرعي في نوع اللصقات العلاجية التي تباع في الأسواق المحلية، وهي تحتوي على كرة صغيرة جدا من الصلب المصفح بالذهب، حيث تعمل هذه الكرة على امتصاص آلام العضلات والمفاصل والعمود الفقري وغير ذلك. فهل يجوز للرجال استعمال هذا النوع من اللصقات العلاجية، كونها تحتوي على طبقة من الذهب؟
هذا السؤال عرض على لجنة الفتوى بوزارة الاوقاف، وقد أجابت اللجنة بالتالي: لا مانع شرعا من استعمال هذه اللاصقات المستفتى عنها لامتصاص آلام العضلات والمفاصل والعمود الفقري. والله تعالى اعلم، وصلى الله على نبينا محمد وعلى آله وصحبه وسلم.

The translation using Google translator

Basqat medication containing gold sheets
What the ruling on the therapeutic plasters sold in the domestic market type, which contains a very small ball of steel laminated with gold, where this ball is working on the absorption of the muscles, joints, spine, and other pains. Is it permissible for men to use this type of therapeutic plasters, they contain a layer of gold?
This question was put to the Fatwa Committee at the Ministry of Awqaf, the Commission has responded thus: I do not mind legitimately use these stickers Poller them to soak up the muscles, joints and spine pain. And God knows, and God bless our Prophet Muhammad and his family and him.

Fig. 2. Economy article with translation using Google translator.

Table 2
Domains and applications of LDA.

Reference	Domain
[7] [8] [9]	Text classification (for English)
[10] [11] [12] [13] [14]	Face recognition
Other LDA based works	Detection and confirmation of counterfeits, Assessing the geographical origin of olive oils, Forensic classification of ballpoint pen inks, Discussing singularity problem, Gender classification, Epileptic magnetoencephalography spikes detection, Estimation event potential source time series, Speaker verification, Electroencephalogram features in dementia patients, Identifying plant part composition

Fig. 2 shows an Economy article; however, it has words that are related to Religion and Health categories that add more challenge to the classification process.

This paper focuses on single label classification; this is why content diversity challenge is clearly observed, however, such challenge might be handled in multilable text classification as an article might assign to more than one category. Based on our investigation of Arabic text documents, we observed more challenges. For instances, we noticed that some Arabic articles have different short stories. That is, a single document contains different short textual segment of completely different topics. Some article also include many English words such as the technology related articles.

3. Literature review

The literature shows that LDA is widely used in pattern recognition problems. Table 2 lists different domains that employed the LDA technique. At first, the table shows some text classification related studies followed by other pattern recognition LDA-based works. The literature shows that face recognition highly employs LDA. The literature also shows that none of the previous text classification studies employs LDA for Arabic text classification. This is one motivation to conduct this research. That is, we aim at demonstrating an original work that uses a machine-learning technique for the Arabic text classification.

The information provided in Table 2 focuses on the LDA method. However, the literature shows that the unsupervised PCA dimensionality reduction technique is also used in different linguistic domains as the following: web page feature selection and classification, text clustering, patent analysis, identify authoritative documents, and language modeling for bag-of-visual words image categorization.

For Arabic, the literature shows that most of the previous works employ feature selection methods that are based on VSM to represent a document. However, VSM represents text as “bag of words” in which each word occurrence or weight

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

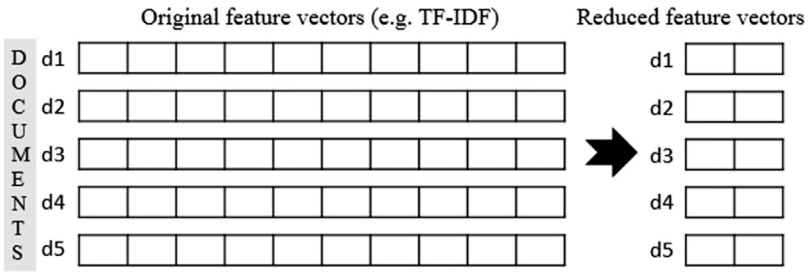

Fig. 3. A demonstration of the transformation process.

corresponds to one independent dimension (i.e. does not preserve the semantic relationships). In contrast, little research has been observed to using features extraction methods such as LDA, PCI, and SVD. Among the little studies, for instances, reference [15] employed the semantic rich SVD method for Arabic text classification. SVD also was used in [16] for Arabic document clustering. Reference [17] used SVD for Arabic text categorization. SVD was also used for Arabic text clustering as indicated in [18].

For recent studies in this field, Reference [19] is a review of the latest research of Arabic text mining. Reference [20] demonstrated a study of the impact of indexing approaches on Arabic text classification. They indicated that the stem is better than full-word form due to smaller data size, which enhances both the processing time and storage utilization as well as scoring higher accuracy. Reference [21] presented an approach for Arabic text categorization based on a graph-based representation. They consider a model in which each document is represented by a graph that encodes relationships between the different named entities.

In fact, this research is trying to highlight the usefulness of LDA for Arabic text classification. The distinguished part of this proposed work is that it did not use any of the existing off-the-shelf toolboxes such as Weka [22], Orange [23], and Rapid Miner [24]. That is, it is not a straightforward to implement feature extraction methods such as LDA in the available machine-learning packages. This is maybe one reason why feature extractions are not widely used for Arabic text classification research, as the researchers generally prefer to use “black boxes” rather than programming from the scratch.

4. Linear discriminant analysis

Employing dimensionality reduction techniques is of importance in the aspects for pattern recognition and text mining. In fact, many studies highlight the benefits of LDA. For instance, it is indicated in [25] that using dimensionality reduction significantly improves the performance. Reference [26] indicated that the dimensionality reduction is the process of finding a suitable lower-dimensional space for several reasons such as: exploring high-dimensional data to discover structure that leads to the formation of statistical hypotheses, visualizing the data using 2-D or 3-D, and analyzing the data using statistical methods, such as clustering or classification. Dimensionality reduction typically employed in very high-dimensional space domains, such as data visualization, data mining, and information retrieval (IR) [27]. Even though using dimensionality reduction is extremely important in many applications, there are some listed constraints [28], such as the complexity of the algorithms for data reduction, predictive / descriptive accuracy (relevant features), and representation of the Data-Mining model (simplicity of representation).

The concept of dimensionality reduction is to find a projection that maximize the discrimination between classes. Fig. 3 demonstrates some of textual feature vectors that reduced using a dimensionality reduction method to lower dimensions. That is, it transforms the five document vectors of ten dimensions to only two dimensions. In this figure, the reduced feature vectors are arbitrary chosen to have two features (i.e. dimensions). In the figure, TF-IDF is the shorthand of Term Frequency – Inverse Document Frequency, the well-known weighting sachem in information retrieval and text mining applications.

For further illustration, Fig. 4 shows 500 randomly generated points of two dimensions each. These vectors are projected on a new space of one dimension. The red points are projected on the green line while the blue cross pints are projected on the magenta line. Hence, the two-dimension vector are now represented with only one dimension with best separation. Even though this is a simple demonstration, nevertheless, LDA can handle feature vectors of large dimensions, such as in face recognition and text classification.

LDA is the process of projecting N data vectors that belong to c different classes into a (c-1) dimensional space in such way that the ratio of between group scatter S_b and within group scatter S_w is maximized [29]. Hence, the classes' centroids in the transformed space spread out as much as possible. Fig. 5 shows the projection of some points (two dimensions) into a new line (z) defined by the set of weight parameters. The discriminant function yields the scores on z is represented using the following formula: $z = w' \; x = w1 \times x1 + w2 \times x2 + \ldots + wk \times xp$, where p is total number of dimensions in each feature vector.

Hence, the first step is to find a set of weight values that maximize the ratio of the between-class scatter to within-class scatter using the training set samples. The eigenvalue decomposition of the between-class (S_b) to within-class (S_w) scatter

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

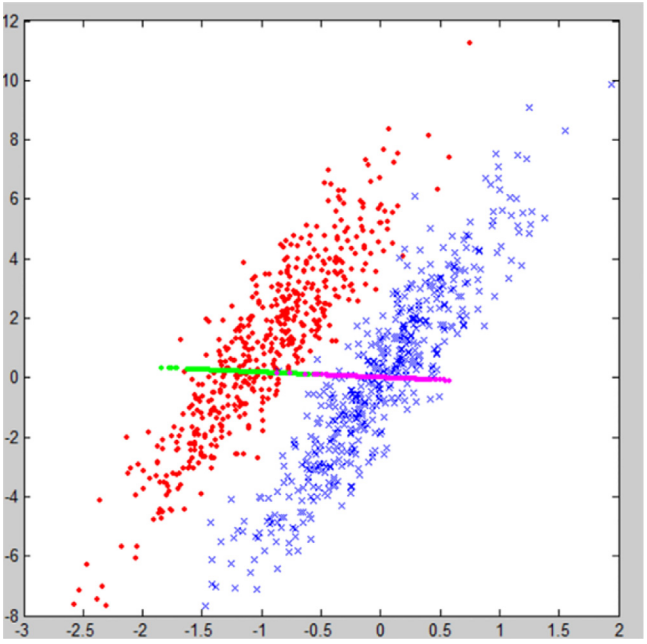

Fig. 4. A projection of two class dataset.

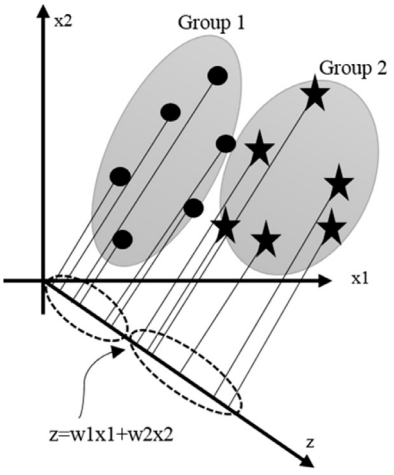

Fig. 5. A projection of two group instances into a new space (z).

matrices is the key towards finding the weight values (eigenvectors). Obtaining the lower dimensions achieved by discarding some eigenvectors that have the smallest corresponding eigenvalues. The steps to obtain the eigenvectors are as follows:

1) Estimate the covariance matrix for each class by: $\sum_j \ (x_j - \mu)(x_j - \mu)^T$ where μ is the mean of the entire dataset and x_j is a sample in the dataset.
2) Estimate within-class (Sw) scatter for each class by: $\sum_{classes\ c} P_c \times covariance_c$ where P_c is the Apriori probability of each class.
3) Estimate the between-class (Sb) scatter by: $\sum_{classes} \ (\mu c - \mu)(\mu c - \mu)^T$ where μ_c is the mean of each class.
4) The between-class (Sb) scatter is also computed by: Sb = entire dataset Covariance-Sw
5) Finding Eigen values and Eigen vectors by: Eigen decomposition ($\frac{Sb}{Sw}$)
6) For two classes case, Eigen values and Eigen vectors are computed by: Inverse (Sw) (m1−m2)
7) The original data vectors are transformed into the new dimensional subspace by: selected maximum Eigenvectors * Original data points
8) The transformed score are compared with classes' centroids using a distance measure.

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

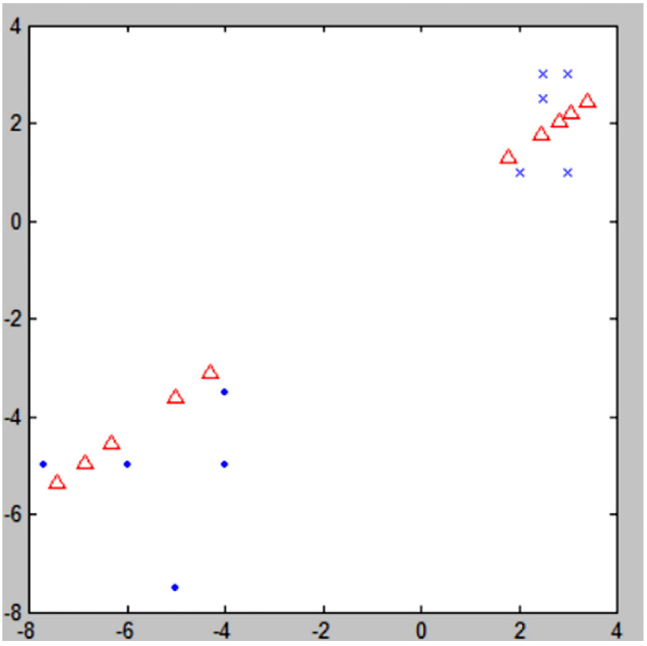

Fig. 6. Two-class data and the projection.

Class 1 (5 instances)	Class 2 (5 instances)
-5.0, -7.5 ➔ 1 -6.0, -5.0 ➔ 1 -7.7, -5.0 ➔ 1 -4.0, -3.5 ➔ 1 -4.0, -5.0 ➔ 1	2.5,2.5 ➔ 2 3.0,1.0 ➔ 2 2.5,3.0 ➔ 2 2.0,1.0 ➔ 2 3.0,3.0 ➔ 2
Apriori probability =P(1)=5/10=0.5	Apriori probability= P(2)=5/10=0.5
Within-class scatter matrix	
Sw=5/10×Cov1+5/10×Cov2 Sw=0.5*[[2.4280 0.2900] + 0.5*[[0.1750 0.1125] = [[1.3015 0.2013] [0.2900 2.0750]] [0.1125 1.0500]] [0.2013 1.5625]]	
Between-class scatter matrix	
$Sb=(m1-m)(m1-m)^T+(m2-m)(m2-m)^T$ m = [-1.37 -1.55], m1 =[-5.34 -5.2], m2=[2.6 2.1] Sb=[[15.76 14.49] + [[15.76 14.49] = [[31.52 28.98]	
[14.49 13.32]] [14.49 13.32]] [28.98 26.64]]	
Computation of the weights using:$Sw^{-1}(m1-m2)$	
w = [[0.7840 -0.1010] * ([[-5.3400] – [[2.6000]) = [[-5.4875] [-0.1010 0.6530]] [-5.2000] [2.1000]] [-3.9652]]	
Transformed values of class 1= 57.17 52.75 62.07 35.82 41.77 Transformed values of class 2= -23.63 -20.42 -25.61 -14.94 -28.35	

Fig. 7. A demonstration of two class Eigendecomposition process.

Fig. 6 shows an example that demonstrates ten data points of two classes. The red points belong to class 1 and the blue cross points belong to class 2. The green triangle shows the projection of all data points. It is clear the projection line maintain the separation between the two classes which facilitates the classification process.

The calculation of the projection in Fig. 6 is demonstrated in Fig. 7. The calculations are performed using Matlab. However, other tools can be used for LDA calculation such as Python.

Once finding the eigenvectors (i.e. the weights) that correspond to some of the maximum eigenvalues, the transformed values of the new space are generated by multiplying the weights with the original data. For classification, the discriminant function z used to classify a new data point. That is, the new sample is transformed to the new space for classification purpose. As indicated, the classification process performed using a distance measure such as Euclidean distance, Cosine similarity measure, etc.

To illustrate the projection process of three-class case, we consider a small data set that contains 15 samples belonging to three classes. The following Fig. 8 shows some information of this set. In this example, we demonstrate how to represent 3-dimenstinal feature vectors into 1-dimenstional feature vectors using LDA.

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

Class 1 (5 instances)	Class 2 (5 instances)	Class 3 (5 instances)
-5,-8,-6 ➔1 -7,-6,-8➔1 -8,-6,-7➔1 -6,-4,-5➔1 -5,-8,-4➔1	1,2,3➔2 3,2,1➔2 2,1,5➔2 2,4,3➔2 1,4,3➔2	10,14,11➔3 12,11,14➔3 11,9,12➔3 14,10,11➔3 12,10,9➔3
Apriori probability=5/15	Apriori probability=5/15	Apriori probability=5/15
The covariance of class 1 (Cov1) [[0.56 -0.36 0.50] [-0.36 0.93 -0.16] [0.50 -0.16 0.83]]	The covariance of class 2 (Cov2) [[0.23 -0.11 -0.16] [-0.11 0.60 -0.16] [-0.16 -0.16 0.66]]	The covariance of class 3 (Cov3) [[0.73 -0.51 -0.05] [-0.51 1.23 0.03] [-0.05 0.033 1.10]]
Within-class scatter: Sw = 5/15×Cov1+5/15×Cov2+5/15×Cov3 [[1.53 -1.00 0.28] [-1.00 2.76 -0.30] [0.28 -0.30 2.60]]	The covariance matrix of the training data points is (C): [[59.40 54.33 56.10] [54.33 55.23 53.21] [56.10 53.21 56.31]]	
Between-class scatter (Sb=C-Sw): [[57.87 55.33 55.81] [55.33 52.47 53.51] [55.81 53.51 53.71]]	Eigenvalues and Eigenvectors: [-92.75 0.0079 -7.0], [[0.78 -0.79 0.053] [-0.25 -0.56 0.68] [-0.55 -0.24 -0.72]]	
After Sorting the Eigenvalues and the Eigenvectors: [0.0079 -7.0 -92.75], [[-0.79 0.053 0.78] [-0.56 0.68 -0.25] [-0.24 -0.72 -0.55]]	The final weight values is the first column of the Eigenvectors since we need just 1 dimension: [[-0.79044526] [-0.56090417] [-0.24613575]]	
The transformed value of the 15 training instances: [[9.91], [10.86], [11.41], [8.21], [9.42], [-2.65], [-3.73], [-3.37], [-4.56], [-3.77], [-18.46], [-19.10], [-16.69], [-19.38], [-17.30]]		
The centroids of the new data representation: class 1, class 2, class 3 [[9.96736178], [-3.61955955], [-18.19096662]]		
The testing set : [[-3,-3,-3],[2,2,2],[10,10,10]] The transformed values: [[4.79], [-3.19], [-15.97]] Clearly, the points belong to class 1, class 2, and class 3, respectively.		

Fig. 8. A demonstration of three class Eigendecomposition process.

Fig. 9 shows the original data in three-dimensional space. The goal of LDA is to have these points represented using less than three dimensions, one or two, of course without losing the discriminative quality for each class.

Fig. 10 shows a chart that represents how LDA transforms the 3-dimensional input feature vectors into a reduced 1-dimensional feature vectors with preserving the differences among the classes' centroids. Even though the training points are clearly separated into three different classes, some overlaps might appear in complex applications. Naturally, the new instance (i.e. the instance in question or the instance that is presented for the classification purpose) is assigned to the closest centroid in the projected space.

Although a single discriminant function z can separate samples into several classes, multiple discriminant analysis is also used to construct a separate discriminant function for each class. That is, instead of solving for the eigenvalue / eigenvector, the linear classification functions can be used for classification. Linear classification functions method estimates the common population covariance matrix by a pooled sample covariance matrix known as (S_{pl}) [1]:

$$S_{pl} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i$$

Where (Si) is the covariance matrix of the class i, k is the total number of the classes, N is the total number of training instances in all classes, and n_i is the number of instances in a particular class. For classification using linear classification functions, assign the new instance (y) to the group for which Li(y) is maximum as follows [1]:

$$L_i(\mathbf{y}) = \bar{\mathbf{y}}_i' \mathbf{S}_{pl}^{-1}\mathbf{y} - \frac{1}{2}\bar{\mathbf{y}}_i' \mathbf{S}_{pl}^{-1}\bar{\mathbf{y}}_i, \qquad i = 1, 2, \ldots, k$$

Where ($\bar{\mathbf{y}}_i$) is the mean of the class i. Fig. 11 shows how to utilize linear classification functions as shown in the previous simple three-class example.

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

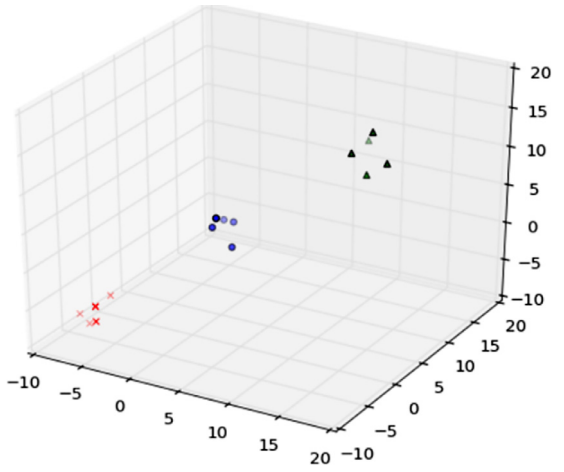

Fig. 9. Three class data set in the original space.

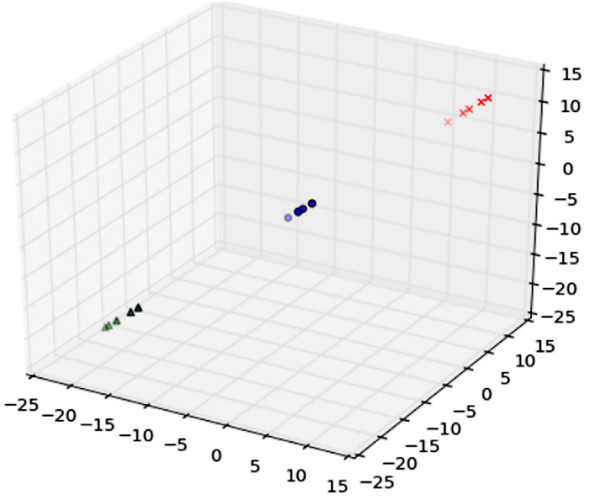

Fig. 10. The data in the transformed space of three classes.

Class 1 (5 instances)	Class 2 (5 instances)	Class 3 (5 instances)
-5,-8,-6 ➔1	1,2,3➔2	10,14,11➔3
-7,-6,-8➔1	3,2,1➔2	12,11,14➔3
-8,-6,-7➔1	2,1,5➔2	11,9,12➔3
-6,-4,-5➔1	2,4,3➔2	14,10,11➔3
-5,-8,-4➔1	1,4,3➔2	12,10,9➔3
mean(class 1)	mean(class 2)	mean(class 3)
[-6.20 -6.40 -6.00]	[1.80 2.60 3.00]	[11.80 10.80 11.40]
S_{pl} =	S_{pl}^{-1} =	Testing set
[[1.53 -1.00 0.28]	[[0.86 0.30 -0.05]	-3 -3 -3
[-1.00 2.76 -0.30]	[0.30 0.47 0.02]	2 2 2
[0.28 -0.30 2.60]]	[-0.05 0.02 0.39]]	10 10 10
The testing set : [[-3,-3,-3],[2,2,2],[10,10,10]]		
The scores of [-3,-3,-3] are { -1.70, -21.48, -224.24}		
The maximum is the first value that means it belong to class 1.		
The scores of [2 ,2 ,2] are { -72.38, 4.24, -95.29}		
The maximum is the first value that means it belong to class 2.		
The scores of [10 ,10, 10] are { -185.46, 45.41, 111.02}		
The maximum is the third value that means it belong to class 3.		

Fig. 11. A demonstration of three class linear classification functions.

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

Table 3
The corpus information.

The training data collection				
#	*Category*	*Number of documents*	*Number of words*	*Number of unique words*
1	Economy	350	225,659	31,942
2	Health	350	151,912	25,835
3	Education	350	224,078	35,294
4	Sports	350	135,473	24,056
5	Tourism	350	192,083	28,218
	Total	**1750**	**929,205**	**80,156***
The testing data collection				
1	Economy	50	22,031	6959
2	Health	50	29,722	9386
3	Education	50	35,338	8961
4	Sports	50	15,168	5482
5	Tourism	50	20,268	6794
	Total	**250**	**122,527**	**24,496***

*It is not algebraic summation since the common words not counted.

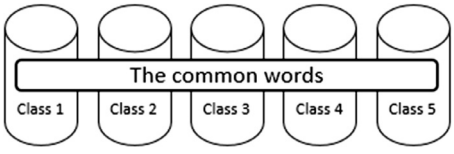

Fig. 12. The common words among the five categories.

5. The proposed method

To investigate the proposed method, we prepared an Arabic text corpus that contains 1750 documents for training and 250 documents for testing. The training set contains 929,205 words with 80,156 unique words. The collected documents belong to five categories as shown in Table 3. The corpus was prepared with help by Alqabas newspaper in Kuwait [30]. Table 3 shows some of the information of the corpus used.

The preprocessing includes deleting all characters that are out of the Arabic alphabetic characters. It also includes deleting numbers, commas, full stops, and all other symbols. A normalization process was also performed to change some Arabic characters, such as (إ→ا) and (أ→ا). The stop word list is declared that contains all common words in the five categories of training documents. Fig. 12 demonstrates what we mean by the common words.

The proposed method is summarized in the following algorithm:

1) Stop word list is declared. In this work, we define the stop word list as the common words in all categories of the training documents. By the common words, we mean the words that appear in all categories (i.e. the five categories).
2) Document Frequency (DF) feature selection threshold is set. Different DF threshold can be investigated to find the optimal performance. For instance, if we use DF as 65, this means that any word that appears in more than 65 different document is appended in the vocabulary.
3) Using the prepared vocabulary, the training and testing feature vectors for all documents are generated using VSM representation.
4) Based on the prepared VSM, LDA is used to generate the transformed feature vectors. The steps to perform LDA process was described in section IV.
5) The centroids of the transformed feature vectors are used for classification based on the Euclidian distance.

6. Experimental results

This section presents the experimental results of the proposed method. The stop word list contains 1846 words. In addition to the stop word list, we discarded all words of one or two characters length since they have no effect in the classification process; an example of single character is the shorthand of doctor (Dr ➔ د). Different DF values were used in the experiments as indicates in Table 4. The DF aims at discarding any word that appears in less than or equal DF threshold. Hence, VSM considered all the words appearing in more than the DF threshold to create 1750 feature vectors of the training set. Each dimension of each created VSM feature vector cantinas the total number of a particular word's occurrences in the document. Of course, the VSM model was also used to create the testing set feature vectors based on the dictionary prepared using the training data set. The total number of features of each feature vector is constraint to the choice of DF threshold. Therefore, the total number of the words in the dictionary is based on the selected DF threshold. Table 4 shows

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

Table 4
The Results using eigenvalue decomposition and linear classification functions.

#	DF	# of original dimensions	Eigenvalue accuracy (%)	Linear functions accuracy (%)
1	64	335	82.0	83.2
2	65	315	81.2	84.4
3	66	305	80.8	82.4

Table 5
The accuracy of each class.

#	Category	Eigenvalue accuracy (%)	Linear functions accuracy (%)
1	Economy	82	88
2	Health	86	86
3	Education	100	98
4	Sports	66	72
5	Tourism	72	78
	Average	**81.2**	**84.4**

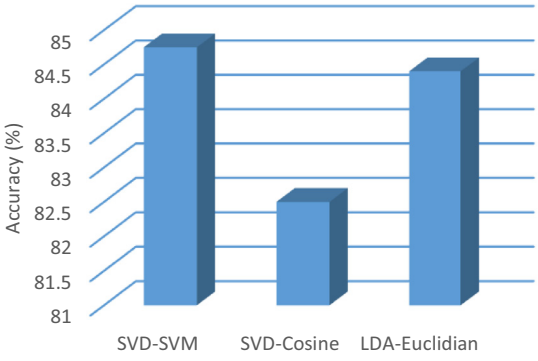

Fig. 13. Performance comparison.

the proposed method performance using both eigenvalue decomposing and linear classification functions. The results show that text classification using linear classification functions outperforms the eigenvalue decomposing method.

Table 5 shows more details of the obtained accuracies for each category using the threshold (DF = 65). The table shows the result for both the eigenvalue decomposition and the linear classification functions methods. Only four LDA dimensions were used for the results presented in Table 5.

We repeated the experiments for different LDA dimensions in the transformed space with the following values: one, two, and three dimensions. The reported accuracies (with DF = 65) are as follows: 46.0%, 63.2%, 74.4%, respectively. We then repeated the experiments using DF equal to 40. This DF gives 923 features; the accuracy scored 70.0% using both eigenvalue decomposition and linear classification functions methods (only four dimensions were used). For comprehensive and precisely evaluation of the LDA performance, we compared the results with our previously published results using the SVD method as demonstrated in [30]. The benchmarking results reported in [30] used almost the same data collection (with a slight change in the number of the categories and the documents).

At that work, we used SVD method with a number of classifiers that include cosine similarity measure, NB, k-NN, NN, RF, SVM, and CT. The best performing classifier was the SVM that scored (84.75%) whereas the cosine similarity measure has less accuracy as (82.5%). In this work, the maximum accuracy is (84.4%) which means that the semantic loss LDA method has almost the same accuracy as the SVD semantic rich method. In addition, this work used the Euclidian distance for classification, which is not known as the best for classification. This gives an indication that LDA is a promising method for Arabic text classification and deserve for more investigation. Fig. 13 shows the performance for SVM, cosine similarity measure, and LDA methods.

Finally, we emphasize that we used the accuracy metric for performance evaluation. Accuracy is simply the ratio of correctly predicted observations to the number of actuals. However, text classification researchers general use other performance metrics such as recall, precision, and F-1 measure. Nevertheless, if the testing set has equal number of documents from each group, then the accuracy measure is enough. In this work, the testing set has equal number of documents as each class has 50 testing documents. This is the reason why we evaluated the performance using only the accuracy metric.

7. Conclusion

In this paper, we have presented the capacity of LDA for Arabic text classification. On other words, the present study is an attempt to understand whether the LDA is adequate for text classification such as other celebrated successful implemen-

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

tations, face recognition is an example. The prior art shows that the LDA is rarely used for Arabic text classification despite its good capabilities in dimensionality reduction. Therefore, this work focuses on the implementation of the LDA method for Arabic text classification as such applications generally contain sizable vocabularies that lead to large features and vectors. The experimental results showed that the performance of the semantic loss LDA feature vectors is almost the same as the semantic rich LSI method. In fact, the literate has little studies to compare machine learning classifiers and dimensionality reduction techniques; nevertheless, our benchmarking comparison showed that the LDA is one of the worthy method as it gives promising results when compared with SVM, k-NN, NN, NB, cosine measure, etc. For instance, the SVM scored accuracy up to 84.75% while the LDA scored 84.4%. This results point out to the important of employing LDA for text classification. In this research, we did not use any of off-the-shelf software platforms such as Weka, as these tools generally do not have all machine-learning options such as extracting features using the LDA technique.

As a future work, it is worthy to investigate other dimensionality reduction techniques such as PCA. That is, PCA might be a good option for Arabic text clustering (i.e. unsupervised learning). It is also worthy to consider other feature selection methods such as Chi-Square Statistic (χ^2), Information Gain (IG), Mutual Information (MI), and Expected Cross-Entropy (ECE). Finally, this paper was designed as tutorial for researchers who are interested in employing the LDA based methods for their own domains as the papers contains some of detailed examples. For recent studies in Arabic text classification, the reader refers to [31] and [32].

Acknowledgment

This work is supported by Kuwait Foundation of Advancement of Science (KFAS), Research grant number P11418EO01. The authors also thank Kuwait University Research Administration for their support.

References

[1] Rencher AC. Methods of multivariate analysis, 492. John Wiley & Sons; 2003.
[2] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41(6):391.
[3] Jolliffe I. Principal component analysis. John Wiley & Sons, Ltd; 2002.
[4] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput 2003;15(6):1373–96.
[5] Avialable: https://translate.google.com/.
[6] Al-Anzi FS, AbuZeina D. Stemming impact on Arabic text categorization performance: a survey. 2015 5th international conference on information & communication technology and accessibility (ICTA). IEEE; 2015.
[7] Torkkola K. Linear discriminant analysis in document classification. IEEE ICDM workshop on text mining; 2001.
[8] Park CH, Park H. A comparison of generalized linear discriminant analysis algorithms. Pattern Recognit 2008;41(3):1083–97.
[9] Li T, Shenghuo Z, Mitsunori O. Using discriminant analysis for multi-class classification: an experimental investigation. Knowl Inf Syst 2006;10(4):453–72.
[10] Martínez AM, Kak AC. Pca versus lda. IEEE Trans Pattern Anal Mach Intell 2001;23(2):228–33.
[11] Liu C, Wechsler H. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans Image Process 2002;11(4):467–76.
[12] Wang X, Tang X. Dual-space linear discriminant analysis for face recognition. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004, 2. IEEE; 2004.
[13] Lu J, Plataniotis KN, Venetsanopoulos AN. Face recognition using kernel direct discriminant analysis algorithms. IEEE Trans Neural Netw 2003;14(1):117–26.
[14] Zheng W-S, Lai J-H, Li SZ. 1D-LDA vs. 2D-LDA: when is vector-based linear discriminant analysis better than matrix-based? Pattern Recognit 2008;41(7):2156–72.
[15] Al-Anzi FS, AbuZeina D. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. J King Saud Univ Comput Inf Sci 2017;29(2):189–95.
[16] Froud H, Lachkar A, Ouatik SA. Arabic text summarization based on latent semantic analysis to enhance Arabic documents clustering. International Journal of Data Mining & Knowledge Management Process (IJDKP); 2013.
[17] Harrag F, Al-Qawasmah E. Improving Arabic text categorization using neural network with SVD. JDIM 2010;8(4):233–9.
[18] Al-Anzi FS, AbuZeina D. Big data categorization for arabic text using latent semantic indexing and clustering. International conference on engineering technologies and big data analytics (ETBDA 2016); 2016.
[19] Alghamdi HM, Selamat A. Arabic web page clustering: a review. J King Saud Univ Comput Inf Sci 2017.
[20] Al-Badarneh A, Al-Shawakfa E, Bani-Ismail B, Al-Rababah K, Shatnawi S. The impact of indexing approaches on Arabic text classification. J Inf Sci 2017;43(2):159–73.
[21] Hadni M, Gouiouez M. Graph based representation for Arabic text categorization. In: Proceedings of the 2nd international conference on big data, cloud and applications. ACM; 2017.
[22] Avialable: http://www.cs.waikato.ac.nz/ml/weka/.
[23] Avialable: http://orange.biolab.si/.
[24] Avialable: https://rapidminer.com/.
[25] Marsland S. Machine learning: an algorithmic perspective. CRC press; 2015.
[26] Martinez WL, Martinez AR, Martinez A, Solka J. Exploratory data analysis with MATLAB. CRC Press; 2010.
[27] Theodoridis S, Koutroumbas K. Pattern Recognition. 4h ed. Academic Press.2010; 2008.
[28] Kantardzic M. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons; 2011.
[29] Duda RO, Hart PE. Pattern classification and scene analysis, 3. New York: Wiley; 1973.
[30] Avialable: http://www.alqabas.com.kw/Default.aspx.
[31] Al-Anzi FS, AbuZeina D. Beyond vector space model for hierarchical Arabic text classification: a Markov chain approach. Inf Process Manage 2018;54(January(1)):105–15 ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2017.10.003.
[32] Al-Anzi FS, AbuZeina D, Hasan S. Utilizing standard deviation in text classification weighting schemes. Int J Innov Comput Inf Control 2017;13(August(4)). http://www.ijicic.org/ijicic-130420.pdf.

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002

Dia AbuZeina Professor AbuZeina received his Ph.D. in Computer Science and Engineering from King Fahd University of Petroleum and Minerals, Saudi Arabia, 2011. He received his M.Sc. in information technology from Southern New Hampshire University, Manchester, USA, 2005. He received his B.Sc. in computer system engineering from Palestine Polytechnic University, 2001. His research interest includes speech recognition and text classification.

Fawaz S. Al-Anzi Professor Al-Anzi received his Ph.D. & M.Sc. in Computer Science from Rensselaer Polytechnic Institute, New York, USA in 1995. He earned his B.Sc. with honors in EE from Kuwait University in 1987. He received the National Research Production Award and Kuwait University Award. He is the founding dean of College of Computing Sciences and Engineering. His research interest includes data science and engineering, text classification and speech recognition.

Please cite this article as: D. AbuZeina, F.S. Al-Anzi, Employing fisher discriminant analysis for Arabic text classification, Computers and Electrical Engineering (2017), https://doi.org/10.1016/j.compeleceng.2017.11.002